<Bioinformatics Center> Pathway Engineering

Bioinformatics Center
- Pathway Engineering -

http://www.bic.kyoto-u.ac.jp/pathway/index.html

Prof
MAMITSUKA, Hiroshi
(D Sc)

Assist Prof
TAKIGAWA, Ichigaku
(D Eng)

PD
WAN, Raymond
(Ph D)

PD
ZHU, Shanfeng
(Ph D)

Scope of Research

With the recent advance of experimental techniques in molecular biology and biochemistry, the research in modern life science is shifting to the comprehensive understanding of a biological mechanism carried out by a variety of biological molecules, including genes, proteins and chemical compounds. The focus of our laboratory is placed on such molecular mechanisms in biological phenomena, represented by biological networks such as gene regulatory networks, metabolic pathways and signal transduction pathways. They are graphs, trees and/or networks in a general computer science terminology. The research objective of our laboratory is to develop computational techniques in computer science and/or statistics to systematically analyze and understand the principles of such biological networks at the cellular and organism level.

Research Activities (Year 2005)

Presentations

Cleaning Microarray Expression Data Using Markov Random Field Based-on Profile Similarity, Wan R, Mamitsuka H and Aoki K F, Twentieth ACM Symposium on Applied Computing, Santa Fe, NM, USA, 14 March.

Efficiently Finding Glycan Motifs Using a Profile Probabilistic Sibling-Dependent Tree Markov Model, Aoki-Kinoshita K F, Ueda N, Mamitsuka H and Kanehisa M, 25th Annual Meeting of the Japanese Society of Carbohydrate Research, Otsu, Japan, 9 July.

Analyzing Metabolic Pathways with Microarray Data Based on Mixtures of Markov Chains, Mamitsuka H, 2005 Japanese Joint Statistical Meeting, Hiroshima, Japan, 13 September.

A Probabilistic Model for Mining Implicit "Chemical Compound - Gene" Relations from Literature, Zhu S, Okuno Y, Tsujimoto G and Mamitsuka H, Fourth European Conference on Computational Biology, Mardid, Spain, 1 October.

A Tree-based Markov Model for Tree-Structure Profiles, Aoki-Kinoshita K F, Ueda N, Mamitsuka H, Goto S and Kanehisa M, Second SIGBIO Meeting, Information Processing Society of Japan, Kyoto, Japan, 7 October.

A Probabilistic Model for Mining Implicit "Chemical Compound - Gene" Relations from Literature, Zhu S, Okuno Y, Tsujimoto G and Mamitsuka H, Second SIGBIO Meeting, Information Processing Society of Japan, Kyoto, Japan, 7 October.

A Profile HMM for Tree Structures to Locate Glycan Structure Profiles, Aoki-Kinoshita K F, Ueda N, Mamitsuka H, Goto S and Kanehisa M, Annual Conference of the Society for Glycobiology, Boston, USA, 10 November.

Grant

Mamitsuka H, Probabilistic Model-based Method for Mining from Structured Data in Bioinformatics, Research Grant from Okawa Foundation for Information and Telecommunications, 1 September 2005 - 30 August 2006.

Mining Biomedical Co-occurrence Data with a Probabilistic Model

Mining literature for biomedical knowledge discovery has become a very active field in bioinformatics recently. One of the important applications is to discover the relationship among genes, proteins, disease phenotype and chemical compounds. Co-occurrence in MEDLINE is a simple and popular technique for discovering possible biological relationships among different entities. This technique is based on the following hypothesis: if biological entity A co-occurs with biological entity B in the same MEDLINE record, A and B should be biologically related with high probability. Here we also employ co-occurrence technique to identify biologically related genes and chemical compounds. We focus on discovering implicit related entities, e.g. "chemical compound - gene", being those which are not in existing co-occurrences in the literature but could be discovered from the co-occurrence data.

We made use of a probabilistic model, which we call a mixture aspect model (MAM), coupled with an efficient algorithm for estimating its parameters. MAM is an extension of a probabilistic model, called the aspect model (AM) developed in natural language processing, with one significant difference of the ability of incorporating different types of co-occurrence data efficiently. A MAM is called *k*MAM when we use *k* different types of co-occurrence data, and 1MAM is equal to AM.

We evaluated our approach by performing experiments on three types of co-occurrence data: gene-gene (GG), compound-compound (CC) and compound-gene (CG) from the MEDLINE records. We extract these data from RefSeq database and corresponding MEDLINE records. In our dataset, we have 22,292 genes and 3,454 chemical compounds. There are altogether 174,077 GG pairs, 20,443 CC pairs and 47,217 CG pairs occurring in 63940 MEDLINE documents.

We evaluated the performance of four different types of MAMs, i.e. AM, 2MAM (+CC), 2MAM (+GG) and 3MAM, using cross-validation on predicting CG pairs. AM uses CG only in training while 2MAM (+CC) uses both CG and CC, and 2MAM~(+GG) uses both CG and GG. 3MAM uses all CG, CC and GG. To examine the effect of the size of the training data set to the performance of the probabilistic model, we set five different ratios of the size of training to test data, 3:1, 2:1, 1:1, 1:2 and 1:3, in the cross-validation experiment. We carried out 50 rounds of this cross-validation to reduce possible biases occurring in only a few rounds and averaged the results obtained. When we add another type of training data, keeping the same training CG pairs for each round of cross-validation, we added one or more other types of co-occurrence data to train 2MAM (+CC), 2MAM (+GG) or 3MAM. Then, the prediction was performed on the same test dataset. We note that AM cannot make any predictions on a CG pair in the test data if one component of this pair does not appear in the training data. Thus, we removed all such co-occurrence pairs in the test data, and the remaining pairs were used as positive test examples. We then randomly generated the same number of CG pairs which are not found in both training and test as negative test examples.

Once we estimated the probability parameters of a probabilistic model from training data, we computed the likelihood of each CG pair in test data and ranked all pairs according to their likelihoods. We evaluated these ranked pairs in AUC (Area Under the ROC curve). Please note that the larger the AUC, the better the performance of the model. We further used the paired sample two-tailed *t*-test to statistically evaluate the performance difference of the two models. Table 1 shows the results.

We also computed the likelihoods of all unknown CG (more specifically, drug-gene) pairs using our approach and selected the top 20 pairs according to the likelihoods. Table 2 shows the 20 pairs. We validated them from biological, medical and pharmaceutical viewpoints.

Model	Ratio of training to test data				
	3:1	2:1	1:1	1:2	1:3
3MAM	**96.0**	**95.5**	**94.5**	**92.8**	**91.5**
2MAM (+CC)	95.0 (**81.4**)	94.5 (**73.9**)	93.2 (**60.3**)	91.1 (**88.6**)	89.6 (**94.9**)
2MAM (+GG)	92.3 (**193.8**)	91.6 (**168.0**)	89.8 (**158.6**)	87.7 (**209.2**)	86.4 (**197.4**)
AM	89.0 (**232.2**)	88.0 (**202.4**)	86.0 (**190.5**)	83.6 (**285.5**)	82.0 (**357.4**)

Table 1. Percentage of the AUCs and the *t*-values (in parentheses) obtained by 50 rounds of cross-validation on compound-gene pairs.

CAS registry number	Drug name	Locus ID	Gene name	Log-likelihood
19545-26-7	Wortmannin	5594	*MAPK1*: Mitogen-activated protein kinase 1	−2.615
16561-29-8	Tetradecanoylphorbol acetate	5590	*PRKCZ*: Protein kinase C, zeta	−2.764
23214-92-8	Doxorubicin	1029	*CDKN2A*: Cyclin-dependent kinase inhibitor 2A	−2.992
73-22-3	Tryptophan	5705	*PSMC5*: Proteasome 26S subunit	−3.000
10102-43-9	Nitric Oxide	959	*TNFSF5*: Tumor necrosis factor, member 5	−3.027
66-81-9	Cycloheximide	5970	*RELA*: V-rel reticuloendotheliosis viral oncogene homolog A	−3.030
33419-42-0	Etoposide	4193	*MDM2*: Transformed 3T3 cell double minute 2	−3.033
50-02-2	Dexamethasone	3458	*IFNG*: Interferon, gamma	−3.037
15663-27-1	Cisplatin	581	*BAX*: BCL2-associated X protein	−3.060
521-18-6	Dihydrotestosterone	2099	*ESR1*: Estrogen receptor 1	−3.061
53-85-0	Dichlororibofuranosylbenzimidazole	2963	*GTF2F2*: General transcription factor IIF, polypeptide 2	−3.103
50-07-7	Mitomycin	7157	*TP53*: Tumor protein p53	−3.104
320-67-2	Azacitidine	6622	*SNCA*: Synuclein, alpha	−3.111
33069-62-4	Paclitaxel	581	*BAX*: BCL2-associated X protein	−3.148
133407-82-6	Leucine aldehyde	7124	*TNF*: Tumor necrosis factor, member 2	−3.203
10540-29-1	Tamoxifen	5241	*PGR*: Progesterone receptor	−3.208
7722-84-1	Hydrogen peroxide	596	*BCL2*: B-cell CLL/lymphoma 2	−3.213
67526-95-8	Thapsagargin	5580	*PRKCD*: Protein kinase C, delta	−3.215
59-14-3	Bromodeoxyuridine	1027	*CDKN1B*: Cyclin-dependent kinase inhibitor 1B	−3.221

Table 2. Top 20 pairs of drugs and genes.